Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph

Maksym Gabielkov, Ashwin Rao, Arnaud Legout

Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph

Maksym Gabielkov
Inria
Sophia Antipolis, France
maksym.gabielkov@inria.fr

Ashwin Rao
Inria
Sophia Antipolis, France
ashwin.rao@inria.fr

Arnaud Legout
Inria
Sophia Antipolis, France
arnaud.legout@inria.fr

ABSTRACT

Twitter is one of the largest social networks using exclusively directed links among accounts. This makes the Twitter social graph much closer to the social graph supporting real life communications than, for instance, Facebook. Therefore, understanding the structure of the Twitter social graph is interesting not only for computer scientists, but also for researchers in other fields, such as sociologists. However, little is known about how the information propagation in Twitter is constrained by its inner structure.

In this paper, we present an in-depth study of the macroscopic structure of the Twitter social graph unveiling the highways on which tweets propagate, the specific user activity associated with each component of this macroscopic structure, and the evolution of this macroscopic structure with time for the past 6 years. For this study, we crawled Twitter to retrieve all accounts and all social relationships (follow links) among accounts; the crawl completed in July 2012 with 505 million accounts interconnected by 23 billion links. Then, we present a methodology to unveil the macroscopic structure of the Twitter social graph. This macroscopic structure consists of 8 components defined by their connectivity characteristics. Each component group users with a specific usage of Twitter. For instance, we identified components gathering together spammers, or celebrities. Finally, we present a method to approximate the macroscopic structure of the Twitter social graph in the past, validate this method using old datasets, and discuss the evolution of the macroscopic structure of the Twitter social graph during the past 6 years.

Categories and Subject Descriptors

H.3.5 [**On-line Information Services**]: Web-based services

Keywords

Twitter; social networks; data mining; graph structure.

SIGMETRICS'14, June 16–20, 2014, Austin, Texas, USA.

ACM 978-1-4503-2789-3/14/06 ...$15.00.
http://dx.doi.org/10.1145/2591971.2591985.

1. INTRODUCTION

Twitter is one of the largest social networks with more than 500 million registered accounts. However, it differs from other large social networks, such as Facebook and Google+, because it uses exclusively arcs among accounts[1]. Therefore, the way information propagates on Twitter is close to how information propagates in real life. Indeed, real life communications are characterized by a high asymmetry between information producers (such as media, celebrities, etc.) and content consumers. Consequently, understanding how information propagates on Twitter has implications beyond computer science.

However, studying information propagation on a large social network is a complex task. Indeed, information propagation is a combination of two phenomena. First, the content of the messages sent on the social network will determine its chance to be relayed. Second, the structure of the social graph will constrain the propagation of messages. In this paper, we specifically focus on how the structure of the Twitter social graph constrains the propagation of information. This problem is important because its answer will unveil the highways used by the flows of information. To achieve this goal, we need to overcome two challenges. First, we need an up-to-date and complete social graph. The most recent publicly available Twitter datasets are from 2009 [16, 9], at that time Twitter was 10 times smaller than in July 2012. Moreover, these datasets are not exhaustive, thus some subtle properties may not be visible. Second, we need a methodology revealing the underlying social relationships among users, a methodology that scales for hundreds of millions of accounts and tens of billions of arcs. Standard aggregate graph metrics such as degree distribution are of no help because we need to identify the highways of the graph followed by messages. Therefore, we need a methodology to both reduce the social graph and keep its main structure.

In this paper, we overcome these challenges and make the following specific contributions.

1. We collected the entire Twitter social graph, representing 505 million accounts connected with 23 billion arcs. To the best of our knowledge, this is the largest *complete* social graph ever collected.

2. We unveil a macroscopic structure in the Twitter social graph that preserves the highways of information propagation. Our method extends the one of Broder

[1]Arcs—that are directed edges—represent the follow relationship in Twitter. If A follows B, A receives tweets from B, but B will not receive tweets from A, unless B follows A.

et al. [7] and can be applied to any kind of directed social graph.

3. We show that not only the macroscopic structure of the Twitter social graph constrains information propagation, but that each component of the macrostructure corresponds to group of users with a specific usage of Twitter. In particular, we show that regular, abandoned, and malicious accounts are not uniformly spread on the components of the macroscopic structure of the Twitter social graph. This result is important to understand how Twitter is used, where users with a specific usage are, and how to sample Twitter without a significant bias.

4. We present a simple methodology to explore the evolution of the macroscopic structure of Twitter with time, we validate this methodology, and show that old datasets from 2009 do not represent the current structure of the Twitter social graph. We explore this time evolution to understand the changes in the usage of Twitter since its creation.

The remainder of this paper is structured as follows. In Section 2, we present our methodology to crawl Twitter and discuss the dataset we collected. We present and discuss, in Section 3, the notion of macroscopic structure, then we describe a methodology to unveil this macroscopic structure. We present the result of applying this methodology to our dataset in Section 4. In Section 5, we propose a simple approach to estimate the evolution of the macroscopic structure of the graph with time, validate this approach, and discuss the evolution of the Twitter social graph from 2007 to 2012. Finally, we present the related work in Section 6, and conclude in Section 7.

2. MEASURING TWITTER AT SCALE

In this section, we describe the methodology used to crawl the Twitter social graph, some high level characteristics of the dataset, the limitations of our crawl, and the ethical issues.

2.1 Crawling Methodology

In order to collect our dataset, we used the Twitter REST API version 1.0 [3] to crawl the information about user accounts and arcs between users. The main challenge of the crawl is that API requests are rate-limited; an unauthenticated host could make at most 150 requests per hour with that API. However this limit could be overcome by using a whitelisted machine. Twitter used to whitelist the servers of research teams and data-intensive services upon request, this service has been discontinued since February 2011, but existing whitelisted machines could still be used. We used four whitelisted machines to perform our crawl, two machines with a rate limit of 20,000 requests per hour and two with 100,000 requests per hour.

We also implemented and deployed a distributed crawler on 550 machines of PlanetLab [2], doubling the crawling rate compared to whitelisted machines only.

We crawled Twitter by user ID, such numeric IDs are assigned for new accounts sequentially, but with gaps [14]. Therefore, we first determined using a random polling that the largest assigned ID is lower than 800 million, then we divided the range from 1 to 800 million into chunks of 10,000 IDs. We selected an upper bound (800 million) much larger than the largest observed ID to be sure to do not miss any account.

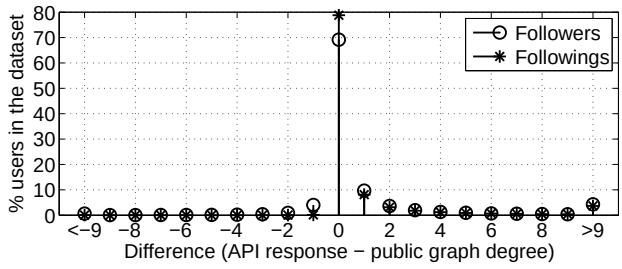

Figure 1: The difference in number of followers and followings between the data from user accounts and the public social graph reconstructed from our dataset.

We performed our crawl from March 20, 2012 to July 24, 2012. We implemented a crawler that assigns chunks of 10,000 IDs to each crawling machine. Then, for a given chunk, each crawling machine performs two steps. First, the machine makes 100 requests for 100 IDs, the maximum number of IDs the lookup method of the API accepts, using an API call [20]. When an ID corresponds to a valid account, we retrieve public numerical, boolean and date information[2]. Second, the machine collects the list of followings for all non-protected and valid accounts with at least one following.

We now define the notions of following, followers, and protected accounts that we use in this paper. Each Twitter account can have *followings* and *followers*. An account receives all published tweets from its *followings*, and all its *followers* receive its tweets. Tweets, and list of followers and followings, are by default visible to everyone. However, users can make their account *protected* which makes this information visible only to its followers. Furthermore, following a *protected* account requires manual approval from its owner [5].

2.2 Limitations of the crawl

There are some accounts that we could not crawl, representing 6.33% of the entire Twitter social graph. We explain in the following the reasons why some accounts are not present in our dataset.

1. 32,112,668 accounts (5.97% of the accounts in our dataset) are protected, so we cannot get their list of followings. The degrees of nodes in the graph we analyzed do not take into account arcs to and from protected accounts.

2. 1,855,945 accounts were referenced in the list of followings of other accounts, but the API lookup did not return any profile information for these referenced accounts. Then we tried to perform further API lookups for these referenced accounts, and we obtained profile information for only 137,899 (7.43%) of them. For the rest, the API lookups did not return any profile information. We guess that these accounts were either

[2]The public information returned by the API call we make is described in this URL `https://dev.twitter.com/docs/platform-objects/users`. We note that the history of the published tweets is not part of it.

deactivated [6] during the crawl or suspended by Twitter because these accounts violated Twitter's terms of use. Users can reactivate their account at any time during 30 days after deactivation, so we guess that the observed 7.43% have reactivated their accounts.

3. For 5,938 accounts, we did not crawl the list of followings because the API consistently returned an error code. We counted the number of followings for such accounts as 0.

4. 1,180 user accounts were lost because our archives with data were partially corrupted due to a system bug on two crawling machines.

The number of followings and followers for each account can be obtained in two ways. Either we get these values from an API call, or we compute them based only on the list of followings for each account. We use the latter to build our social graph, so we cross-validated the number of following and followers using the latter method with the former one. We see in Figure 1 that there is no difference between the numbers of followers (resp. followings) returned by the API and the number of followers (resp. followings) in the social graph we computed for 69.14% (resp. 78.79%) of the collected accounts. The difference observed for the other accounts is due to three different reasons. First, our graph does not include protected accounts and their incoming and outgoing arcs, so the number of following and followers in the computed graph is smaller than from the API, which explains that we observe a higher number of positive differences in Figure 1. Second, there is a delay between the time the account information was crawled and the time the list of followings was crawled because of the implementation of the crawler described in Section 2.1. This delay of 9 hours on average (9.5 minutes median) causes a difference in the number of followings reported by the API and the number of followings obtained by computing the social graph, because some arcs might be added or removed during this delay. Third, we crawled all accounts during a four months period. So a given account crawled at time T might be followed (resp. unfollowed) by accounts after time T, accounts that we crawled after they added (resp. removed) the follow links. Thus, there is a larger (resp. smaller) number of followers for this given account in the computed social graph than returned by the API.

2.3 Measured Twitter Social Graph

We collected all Twitter accounts, consisting of 537 million accounts at the end date of our crawl in July 2012, and accounts' public information (including account creation date, number of published tweets, number of followings and followers, etc.) We remind that there are 5.97% of all accounts (32 million) that are protected, which means one needs their approval to get the lists of their followings. So we collected the list of all followings for non-protected accounts only, resulting in a social graph with 505 million nodes and 23 billion arcs. The average node in-degree of this graph is 45.6, the median is 1, and the 90th percentile is 33.

Our dataset is, to the best of our knowledge, the largest and most complete dataset of a social network available today. We also believe that it will be harder in the future to collect such a large and complete dataset. Indeed, companies are taking measures to prevent large crawls of their social networks. For instance, Twitter is no more whitelisting machines. Moreover it has discontinued on June 11, 2013 the API 1.0 that supported anonymous requests and use of already whitelisted machines. The new API 1.1 requires user authentication for each request making crawls harder and longer to perform. For these reasons, we acknowledge that our dataset has value to communities interested in social graphs, and we release it for academic use only (with precautions described in Section 2.4) [1].

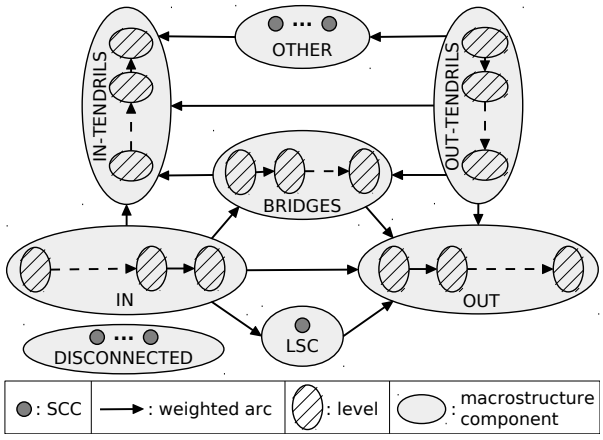

Figure 2: Macrostructure of any directed graph.

2.4 Ethical Issues

There are two main ethical issues with large scale measurement studies. First, we need to take care of users privacy. All data collected in this study are publicly available through the Twitter API, the Twitter applications, and the Twitter Web site. In particular, we did not collect any data that is not publicly available, or did not work around any protection mechanisms.

Second, we need to respect Twitter terms of use. We used the regular Twitter API to perform our crawl. We made half of our crawl using machines whitelisted by Twitter, and half of the crawl using a distributed crawler which used the regular Twitter API and conformed to its rate constraint. On average, we generated from the distributed crawler around 20 requests per second to the API, a rate of requests we believe to be negligible for the Twitter infrastructure.

We release our dataset [1] that consists of the Twitter social graph in the format of an adjacency list. In the released dataset each account ID is anonymized.

3. GRAPH ANALYSIS METHODOLOGY

We start discussing the motivation and insights behind the analysis of the macroscopic structure—henceforth called the macrostructure—of the Twitter social graph. There is a fundamental difference between directed social graphs such as Twitter and other directed graphs such as the Web. In a directed social graph, not only the links among accounts show the influence of accounts, but they also constrain the propagation of information. Therefore, unveiling the macrostructure of a social graph sheds light on the highways of information propagation.

However, it is a challenge to extract a macrostructure on a social graph of the size of Twitter. The intuition behind

our macrostructure analysis is the following. We want to understand how the Twitter graph constrains the flow of information. Therefore, we start by identifying all the strongly connected components (SCCs) that are components with a directed path between any two nodes. In such components, the information can freely circulate, so we abstract each of these components by a single node. After this stage, we obtain a directed acyclic graph (DAG) that is half of the size of the original graph (in terms of number of nodes), still too large to be analyzed. Consequently, the next stage is to group nodes in this DAG based on their connectivity to the largest SCC. As discussed in the following, the largest SCC represents roughly half of the nodes. This is large and there is undoubtedly an interesting analysis to make on this component, but we keep this analysis for future work and focus in this paper on the macrostructure. After this stage, we have 8 components representing a tractable graph. We now describe the details of this process.

We compute the macrostructure of the Twitter social graph in two stages. In the first stage, we use the Tarjan algorithm [23] to compute the SCCs of the Twitter social graph. Then, we replace each SCC with a single vertex, and the multiple arcs between any two vertices with a weighted arc of weight equal to the number of arcs it replaces. As a result, we obtain a directed acyclic graph.

In the second stage, to uncover the macrostructure of the directed acyclic graph shown in Figure 2, we use the following procedure. We first identify the Largest Strongly Connected (LSC) component, the component with the largest number of original nodes. From this LSC component, we run a breadth first search (BFS). We define the set of vertices we find to be the OUT component, that is the set of nodes with a directed path from the LSC component. Inside the OUT component we distinguish *levels* (shown as hatched ellipses on Figure 2). Each level is a bin of SCCs that have the same distance from the LSC component. Then we run a reverse BFS from the LSC component and define the set of vertices we find to be the IN component which is a set of nodes with a directed path to the LSC component. Similarly to OUT we distinguish levels inside the IN component based on the distance to the LSC component. Next, we perform a BFS starting from the IN component and a reverse BFS from the OUT component, reachable nodes that were not yet in the LSC, IN or OUT components were identified as IN-TENDRILS and OUT-TENDRILS respectively. Inside the tendrils we can also identify levels depending on the distance to the components these tendrils are growing from. We separated nodes that were identified as both IN-TENDRILS and OUT-TENDRILS into the BRIDGES category that consist of accounts connecting the IN and OUT bypassing the LSC component, we can also distinguish levels based on the distance to OUT and distance to IN. After that we put the nodes that were not categorized on previous steps into the OTHER category when there is an undirected path from them to categorized nodes or to DISCONNECTED category otherwise. All the possible arcs between the components of the macrostructure are shown on Figure 2.

The methodology we describe and the macrostructure representation is inspired from the work of Broder *et al.* [7] in the context of the Web for 203 million Web pages. However, our methodology is significantly different from the one presented by Broder *et al.* Indeed, unlike our methodology that is exhaustive, they used a small random sample

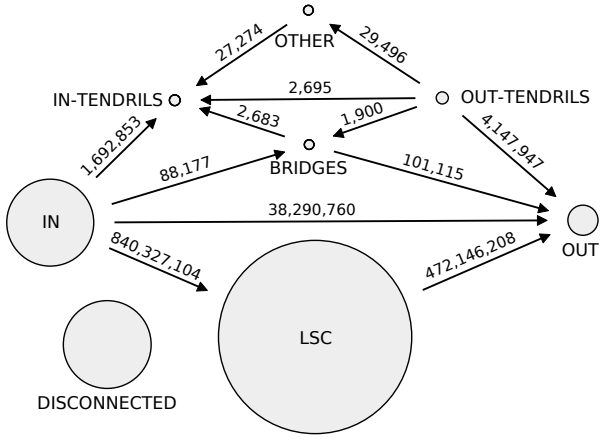

Figure 3: Macrostructure of Twitter in July 2012. *The size of the circles is proportional to the number of accounts in components. The labels on arrows give the number of arcs between components.*

of 570 nodes from the LSC component to find other components. This difference in methodology has two important consequences. First, we perform a complete and accurate classification of all accounts, which is not possible with the methodology of Broder *et al.*, a methodology only intended to show the macrostructure, but not to accurately classify accounts. Second, the macrostructure we describe is more detailed and accurate. In particular, unlike Broder *et al.*, we identified a new component called OTHER, the structure of levels within components, links between components, and the exact number of such links.

In addition, insight we can get from unveiling the macrostructure of the Web is very different from the insight we can get from unveiling the one of a directed social graph such as Twitter. Indeed, the Web forms a directed graph and the arcs among Web pages are hypertext links. Therefore, the directed graph of the Web represents the paths to access Web pages, but no information propagates along the arcs of the graph. On the contrary, the directed graph of Twitter consists of the follow relationship among accounts. Each tweet published can only propagate along the paths of this graph. Therefore, whereas the notion of content propagation is irrelevant in the context of the Web graph, it is central in the context of the Twitter graph.

In summary, we present a method to compute the macrostructure of any directed graph. Figure 2 is not specific to Twitter and can be applied to any directed graph, and in particular to social graphs, where the components group together accounts with different roles in the social graph. This representation is, to the best of our knowledge, the first attempt to extract *exhaustively* a macrostructure of a large social graph, such as the one of Twitter, taking into account the connectivity of accounts in this graph. In Section 4, we will discuss the role of the Twitter accounts, depending on the component they belong too.

4. THE MACROSTRUCTURE OF TWITTER IN JULY 2012

Exploring the macrostructure of the Twitter social graph is interesting because it sheds light on how information prop-

Component	Top followed (%)	Top following (%)	Top tweeting (%)	Experts (%)	Verified (%)	Suspended (%)
LSC	96.95	100	88.66	94.28	97.01	1.17
OUT	3.05	0	10.79	1.33	2.99	0.43
IN	0	0	0.07	0.01	0	1.77
DISC.	0	0	0.47	0.01	0	5.11
OUT-T.	0	0	0	0	0	0.18
IN-T.	0	0	0	0	0	0.49
BRID.	0	0	0	0	0	0
OTHER	0	0	0.01	0	0	1.25

Table 1: Distribution of noteworthy accounts among components. *The first three columns represent the 10,000 accounts with the largest number of followers, followings, and tweets for the entire Twitter social graph. The fourth column represents the 2.91 million experts identified by Sharma et al. [22] as influential users in their field (the sum of this column is not 100% because 4.37% of the experts are not present in our dataset, most likely because they closed their account, or have been suspended). The fifth column represents the accounts verified by Twitter. The last column represents the percentage of suspended accounts.*

agation is constrained. However, this macrostructure would be even more interesting if we can map specific usages of Twitter to components in this macrostructure. Unraveling a correlation between accounts usages and the macrostructure will improve the understanding of how Twitter is used.

In this section, we dissect the macrostructure of the Twitter social graph, focusing on regular, abandoned, and suspicious accounts. i) **Regular accounts** are by definition accounts that are neither abandoned or suspended. Such accounts show the regular activity on Twitter. ii) **Abandoned accounts** are accounts with few followers and followings, and no recent tweet activity. Such accounts are important to understand Twitter adoption and to accurately quantify the bias when analyzing Twitter, bias due to these accounts that do not take part in any social activity. iii) **Suspicious accounts** are often suspended by Twitter because they infringed its terms of use. We checked that most suspended accounts show evident signs of malicious activity (bunch of sequentially generated accounts, accounts' user name generated with automatic patterns, etc.). There is no ground truth for the malicious activity, but the notion of suspended accounts is a reasonable metric to detect (in retrospect) malicious accounts [24]. For the purposes of our study we have recrawled a set of 1 million random users from our dataset on May 6, 2013 to check if they are still active. In the rest of the paper, we refer to the number of suspended accounts as the number of accounts for which Twitter returned the 'suspended' status during this recrawl.

Figure 3 shows the macrostructure of Twitter computed with the methodology presented in Section 3. We identify 8 components in this graph, with 4 of them (LSC, OUT, IN, DISCONNECTED) representing 98.96% of all Twitter accounts; so we focus on them.

	Arcs (%)		Tweets (%)	Accounts (%)
	followers	followings		
LSC	98.01	96.13	98.05	50.71
OUT	1.96	0.02	1.49	5.30
IN	0.02	3.83	0.25	21.36
DISC.	<0.01	<0.01	0.21	21.60
Others	<0.01	0.02	<0.01	1.03
Total	23×10^9		127×10^9	505×10^6

Table 2: Distribution of the arcs, tweets and accounts per component. *At the scale of the entire Twitter social graph, there is the same number of followings and followers, because they represent the same notion of arc. But, for each component, the number of followings and followers might be different due to the ingress and egress arcs, so we make a distinction between followings and followers for each component.*

Component	No follower (%)	No following (%)	No tweet (%)
LSC	0	0	23.87
OUT	0	92.97	61.82
IN	96.13	0	60.10
DISCONNECTED	99.63	99.63	79.31
OUT-TENDRILS	99.13	0	73.20
IN-TENDRILS	0	98.78	70.40
BRIDGES	0	0	67.34
OTHER	51.39	46.67	67.56

Table 3: Percentage of accounts with no follower, no following or no tweet per component.

The LSC (Largest Strongly Connected) component is the core of the regular Twitter activity. Indeed, according to Table 1, the LSC component contains 96.95% of the 10,000 most followed accounts, 100% of the 10,000 accounts that follow the most, 88.66% of the 10,000 accounts that tweet the most, 94.28% of the 2.91 million experts identified by Sharma *et al.* [22] as influential accounts in their field, and 97.01% of the verified accounts [4] that are accounts of highly sought users (in music, acting, politics, etc.) that Twitter verified to be authentic. In addition, Table 2 shows that more than 96% of the following and follower links, and 98.05% of the tweets are for accounts in the LSC.

However, it is wrong to believe that the LSC component is the only one that matters when studying Twitter, other components contain a lot of accounts with specific roles in the Twitter ecosystem. We see in Table 2 that the LSC contains only 50.71% of all accounts. This is surprising because it is easy to be part of the LSC component, an account only needs one following and one follower already in the LSC component. Also, we observe that a large fraction of the suspicious activity in Twitter is outside of the LSC component, as we see in Table 1 (last column). Finally, when looking at the percentage of accounts with no follower, no following, or no tweet, we see in Table 3 that each of the four main components has fundamentally different characteristics. Indeed 92.97% of the accounts in the OUT component have no following, 96.13% of the accounts in the IN component have no follower, and almost all accounts in the DISCONNECTED

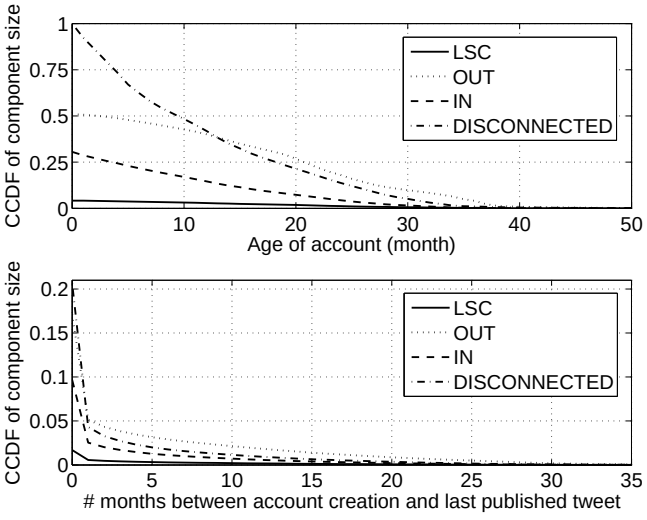

Figure 4: Characterization of abandoned accounts. *(top) Identification of old abandoned accounts. CCDF of accounts with at most one follower and one following in a component according to the account creation date. (bottom) Characterization of accounts who published at least one tweet. CCDF of the duration between the creation date of an account and the date of its last published tweet for accounts with at most one follower and one following.*

component have no following and no follower. Moreover, at least 60% of the accounts in these three components never sent any tweet, whereas it is only 23.87% for the LSC.

In summary, we see that even if most of the regular Twitter activity is in the LSC component, other components contain half of the Twitter accounts and present characteristics worth studying. In the following, we dig into each component to discuss its main characteristics.

4.1 LSC Component

We have seen that most of the regular Twitter activity is in the LSC component. However, due to the simplicity to belong to the LSC component, many abandoned and suspicious accounts also belong to it.

4.1.1 Abandoned Accounts

Most accounts with one following and one follower in the LSC are abandoned accounts. We see in Figure 4 (top, solid line) that there are 4.18% of accounts in the LSC component with one following and one follower. In addition, out of the accounts with one following and one follower in the LSC component, 86.34% are more than 6 months old and 59.57% never sent any tweet.

In summary, a large fraction of accounts in the LSC component with one following and one follower did not have any change in their number of followings and followers for months and did not send tweets recently. Considering that it is unlikely that such accounts will actively follow a single other account for month (so no serious follow activity) without tweeting anything (so no publishing activity), it is reasonable to believe that these accounts are abandoned.

Component	Top followed (%)	Top following (%)	Top tweeting (%)	Top following with ≤ 1 follower (%)	Top tweet with ≤ 1 follower (%)
LSC	0.33	1.15	1.99	97.83	3.02
OUT	1.15	10.30	5.20	0.45	5.26
IN	2.78	96.87	3.87	96.87	3.89
DISC.	1.38	1.33	7.43	2.84	7.48

Table 4: Percentage of suspended accounts (on the 6th of May 2013) per component for 5 outlier categories. *The first three columns represent the 10,000 accounts with the largest number of followers, followings, and tweets for the entire Twitter social graph. The fourth column is for the 10,000 accounts with the largest number of followings and at most one follower. The last column is for the 10,000 accounts with the largest number of tweets and at most one follower.*

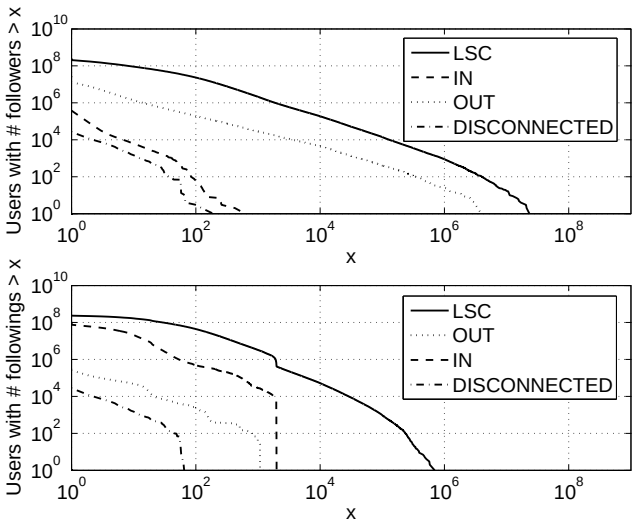

Figure 5: Distribution of followers (top) and followings (bottom) by category. *Accounts with no follower (top) and no following (bottom) are filtered out (see Table 3)*

4.1.2 Suspicious Accounts

The LSC component also contains suspicious accounts. We present in Table 4 the percentage of suspended accounts per component for five outlier categories. An *outlier* account is followed, following, or tweeting much more than a regular account, thus it is a good candidate for suspicious activity. The first three columns represent accounts with the largest number of followers, largest number of followings, and largest number of tweets. The fourth and fifth columns are for the accounts with the largest number of followings and tweets, but with at most one follower. We consider this notion of outliers because following a lot of accounts is a known technique used by spammers [24]. To reduce the impact of spammers, we remind that Twitter imposes a limit of 2,000 followings for accounts with no follower, and then a

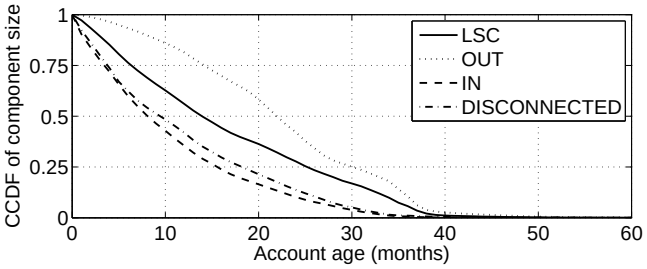

Figure 6: Age of accounts in each component. *CCDF of accounts in a component according to the account creation date.*

linear increase with the number of followers. Accounts close to this limit and with at most one follower are more likely to be spammers. The last column is for accounts that send the largest number of tweets, but with at most one follower. This is also a suspicious behavior, because it is strange to send a lot of tweets if nobody (or a single other account) follows them. Spammers can send a lot of tweets to interfere with trending topics or the Twitter search functionality, and to direct messages to a specific user using *@mentions* [24].

Considering the huge number of suspicious accounts, we cannot afford to manually inspect all of them. Therefore, we consider a suspicious account to be malicious if it was suspended by Twitter, see Table 4.

As expected, the top followed accounts in the LSC component are regular, only 0.33% have been suspended. Indeed, it is complex to manipulate the number of followers, because it requires to either manipulate other accounts in order to incite them to follow, or to create fake accounts whose only one goal is to follow. More surprising, the top following accounts are also regular for Twitter, only 1.15% have been suspended. We expect accounts that follow a lot of other accounts to be spammers, but, according to Figure 5 (bottom), the LSC component is the only one to have accounts that break the limit of 2,000 followings. So the top following in the LSC component also have a lot of followers, thus the low number of suspended accounts.

Then we observe in Table 4 two important behaviors that characterize well the outlier activity in the LSC component. First, 97.83% of the top following with at most 1 follower have been suspended. This means that most of the accounts in the LSC component close to the limit of 2,000 followings are malicious. Second, only 3.02% of the top tweeting accounts, but with at most a single follower have been suspended. The rest looks like regular for Twitter. By manually inspecting these accounts that looks regular for Twitter, we found bots used as an interface to job forums, news site, Yahoo!Answers, YouTube published videos, etc. So, it seems that Twitter is used by developers to generate a stream of data collected from third party Web sites. As these accounts have only one follower, we guess that they are either used for tests only, or that the developers are using a Twitter widget to embed their account timeline into a Web site.

4.2 OUT Component

The OUT component represents all Twitter accounts with a directed path from the LSC component. In addition, these accounts can also have directed paths from other components, but no account in OUT can have a directed path to any other component (directed paths among OUT accounts are possible, so if an OUT account has following links, they necessarily come to other OUT accounts).

4.2.1 Regular Accounts

A specificity of the OUT component is that a small set of accounts (belonging to celebrities) attract most of the follower links for this component. These are regular OUT accounts. We see in Figure 3 that more than 500 million links between components are directed to OUT, 37.93% of all inter-components links, whereas the OUT component represents only 5.30% of all accounts. Also, we see in Table 2 that accounts in OUT presents 1.96% of all follower links, which make it the second component with the largest number of follower links (we sum all follower links for all accounts in a given component). Among the 100 accounts that have the largest number of followers, we found that there are 35 verified accounts representing 12% of the arcs from the LSC to OUT. These accounts are owned by celebrities that belong to the OUT component because they do not follow any other account.

We observe another interesting specificity of the OUT component in Figures 4 (top) and Figures 6. The OUT component is the only one to show an inflection point for both curves around 20 months, meaning that the proportion of recent accounts in the component is lower than for other components. To explain this inflection point, we need to characterize the kind of accounts that stay in the OUT component. According to Table 3, 92,97% of the OUT accounts have no followings, but they all have at least one follower because they belong to the OUT component. These accounts are what we call selfish (they are not interested in tweets from other accounts), a decreasing trend in Twitter in the past two years. We will discuss further this trend in Section 5.3.

4.2.2 Abandoned Accounts

As we discussed in Section 4.1, most accounts with at most one following and one follower are also abandoned accounts for the OUT component. We see in Figure 4 (top) that 50.94% of the accounts have at most one following and one follower, and 40.11% are more than 1 year old. We see in Figure 4 (bottom) that 82.89% of the accounts with at most one following and one follower never sent a tweet, and that only 5.13% of the accounts with at least one following and one follower sent a tweet more that 1 month after their creation date. This is a consequence of the *Find friends* feature available in Twitter that allows users to search their entire contact lists for Twitter accounts. By default, once the search is done, all accounts are checked to be followed. As a consequence, we observe many accounts in the LSC component that followed abandoned accounts in the DISCONNECTED component, making these abandoned accounts move to the OUT component.

4.2.3 Suspicious Accounts

There are fewer malicious accounts in the OUT component than for other components. We see in Table 4 that the percentage of suspended accounts for outlier accounts is low for the OUT component. We explain the low number of suspended accounts for the top followings because no account reaches the limit of 2,000 followings, and that few accounts have more than a hundred followings, see Figure 5

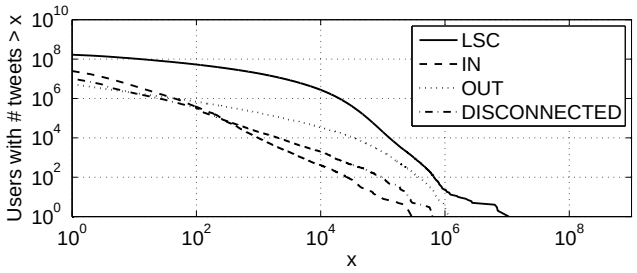

Figure 7: The distribution of number of tweets by component. *Accounts with no tweets are filtered out (see Table 3).*

(bottom). As long as an account in OUT follows an account in the LSC, it belongs to the LSC, so spammers using following links to spam are likely to escape in the LSC component. We explain the low number of suspended accounts for the top tweeting with at most one follower because (as for the LSC component) most of these accounts are operated by bots. Finally, we see in Table 1 that out of the 4 main components, OUT is the component with the smallest number of malicious accounts.

4.3 IN Component

The IN component is much different from the two previous ones because accounts in this component have few followers (see Figure 5, top) and the distribution of the number of tweets is very different (see Figure 7) from the ones of the LSC and OUT components. The IN component contains the second largest fraction of abandoned and suspicious accounts, after the DISCONNECTED component.

4.3.1 Regular Accounts

The regular users for the IN component are passive followers, that are accounts who follow accounts in the LSC, but never publish tweets and are not followed. Indeed, in Tables 2 and 3 we see that the IN component contains 21.36% of all Twitter accounts, but 96.13% of them have no follower, and 60.10% of them published no tweet (we remind that accounts with followers in IN are followed by other accounts in IN only). This component consists of accounts who follow accounts in the LSC (99.6%) or an account in IN (0.4%). We will see in Section 5.3 that the trend of accounts to be passive followers on Twitter (that is, belong to IN component) has been growing since 2009.

Many accounts belonging to the IN component move to the LSC component. We see in Figure 4 (top) that 30.56% of the accounts in the IN component have at most one following and one follower, but that only 14.61% are more than one year old. So even if few of them have been tweeting close to the creation date of their accounts (see Figure 4, bottom), it is likely that they moved to the LSC component and tweeted from it. Indeed, we see in Table 1 that only 1.77% of the accounts in the IN component have been suspended, but that accounts are much younger in the IN component than in the LSC and OUT components, see Figure 6.

4.3.2 Abandoned Accounts

Whereas 96.13% of the accounts in the IN component never published any tweet (see Table 3), the fraction of abandoned accounts is much lower in this component than in the OUT and DISCONNECTED components. Indeed, we see in Figure 4 (top) that only 30.56% of the accounts in the IN component have at most one following and one follower, and 20.88% have at most one following, one follower, and never published any tweet. Moreover, according to Figure 5 (bottom), 23.04% of the accounts follow at least 10 other accounts, thus a passive follower activity.

4.3.3 Suspicious Accounts

The IN component contains many malicious accounts among the outliers. We see in Table 4 that 96.87% of the accounts with the largest number of followings are suspended. We note that all top followings have at most 1 follower in this component. There is also 3.87% of the accounts that tweeted the most that were suspended. For the rest, after manual inspection, we also found, as for the two previous components, that they are used by bots.

Finally, the IN component has a very interesting property for people looking for a reliable metric to assess influencers. Cha *et al.* [9] show that the number of followers is not a reliable metric, because users perform link farming [10] to increase their number of followers. However, this is a rare problem in the IN component. Indeed, accounts in the IN are clearly not interested in increasing their number of followers (see Figure 5, top) thus the accounts they follow will not be biased by this problem. Evaluating the benefit of considering accounts in the IN to assess influencers is an interesting problem for future work.

4.4 DISCONNECTED Component

Accounts in the DISCONNECTED component, like in the IN one, have few followers (see Figure 5, top) and the distribution of their number of tweets is very different (see Figure 7) from the ones of the LSC and OUT. The DISCONNECTED component contains the largest fraction of abandoned and suspicious accounts. There are almost no regular users in this component.

4.4.1 Abandoned Accounts

A specificity of the DISCONNECTED component is that is contains a lot of abandoned accounts. In spite of being the second largest component with 21.6% of all accounts (see Table 2), 78.94% of accounts in the DISCONNECTED component have no followers and no followings, and never published any tweet. Furthermore, 72.44% of accounts in DISCONNECTED component are older than one month. Therefore, we can conclude that the DISCONNECTED component has, by far, the largest number of abandoned accounts. We see in Fig. 4 (top) that 99.97% of its accounts have at most one following and one follower, but only 41.93% of them are older than 12 months. Like for the IN component, many account in the DISCONNECTED component are recent (see Figure 6), thus some accounts in this component have moved to another component.

4.4.2 Suspicious Accounts

Finally, we see in Table 1 that the DISCONNECTED component contains the largest fraction of malicious accounts, but we don't observe in Table 4 an outlier category grouping them. Indeed, most accounts have no followings, no followers and no tweets, so the number of outliers is much smaller than our sample size.

In summary, the DISCONNECTED component hosts a

lot of abandoned accounts and a large fraction of the malicious activity on Twitter, it is also a transitional place for new accounts before they migrate to another component.

4.5 Other Components

The smallest components, IN-TENDRILS, OUT-TENDRILS, BRIDGES, and OTHER represent 1.03% of all accounts. Most accounts in these components are either accounts created for test, or new accounts that will migrate to another component after some time. We do not discuss deeper these components as their impact on the Twitter social graph is small compared to the 4 main components.

4.6 Discussion

We can draw several important lessons from the results discussed in this section.

First, the macrostructure of the Twitter social graph significantly constrains the propagation of information. Therefore, models of information propagation in social networks might lead to wrong results when abstracting the underlying social graph. This work sheds light on how to correctly abstract the social graph, and because the macrostructure is reasonably simple, with 3 main components with active accounts, we believe it is possible to model the underlying graph constraint.

Second, we identify a correlation between components in the macrostructure and the usage of accounts in these components. This result challenges the sampling techniques that follow arcs (such as random walks or bi-directional breadth first search) because the statistical validity of the sample might be low. For instance, all sampling techniques following arcs that start from well connected (or active) accounts will miss all of the malicious activity located in the DISCONNECTED component.

Last, the identification of the role of accounts in each components is important to understand who are the influencers in Twitter. For instance, as discussed in Section 4.3.3, users try to increase their popularity in Twitter by either offering reciprocation to the users that accept to follow them, or by buying follower links on the black market. Therefore, we can identify real influencers by focusing exclusively on the followers in the IN component, removing suspicious accounts by filtering out all accounts younger than, e.g., six months.

5. EVOLUTION OF THE MACROSTRUCTURE OF THE TWITTER SOCIAL GRAPH WITH TIME

In this section, we discuss the evolution of the macrostructure of the Twitter social graph with time from January 2007 to July 2012. To present this evolution, we first describe the estimation technique we use to estimate the Twitter social graph in the past. Then we validate our technique using two public datasets collected in 2009 [16, 9]. We discuss the evolution of the macrostructure of the Twitter social graph with time and explain how new accounts led to the time evolution we observed, shedding light on the evolution of the usage of Twitter in the past 6 years.

5.1 Methodology to Estimate the Macrostructure

The evolution with time of the macrostructure of the Twitter social graph is interesting, because it shows the evolution of the Twitter usage. We have seen in Section 4 that components represent specific categories of usage. However, the Twitter API does not give access to the past social graph of Twitter.

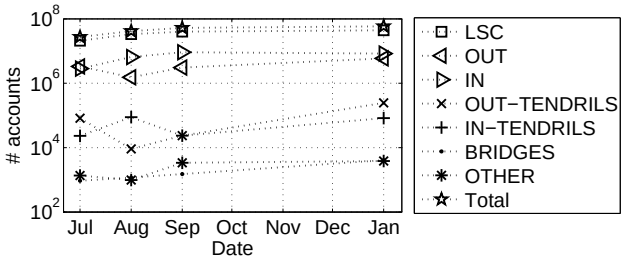

Figure 8: Comparison of our estimated graphs of 2009 (labeled Jul and Jan) with two existing Twitter datasets made in August [16] and September [9] 2009. *Our simple methodology gives an approximation of the macrostructure of the Twitter social graph that is consistent with existing datasets.*

We propose a simple approach to approximate the macrostructure of the Twitter social graph. The dataset we describe in Section 2.3 covers all Twitter accounts in July 2012 (with the limitation described in Section 2.2), and for each account we have the creation date. To approximate the macrostructure of the Twitter social graph at date D, we remove from our dataset all accounts created after this date, and all arcs to and from these accounts. Then, we use the methodology described in Section 3 to compute the macrostructure of the resulting graph at date D.

This simple methodology has two important limitations. First, we do not have any suspended and deactivated accounts in our dataset. Accounts are suspended by Twitter because they infringed the terms of use, most of the time they are spammers. Deactivated accounts have been closed by users themselves. We believe such accounts, when they were still active, had a small impact of the Twitter social graph. Second, the Twitter API does not give access to the arc creation date[3]. Therefore, we assume that all arcs between any two accounts in July 2012 existed at date D as long as the two accounts existed at this date; equivalently we assume that if there is an arc between two accounts, it is created close to the creation date of the youngest account. We are aware that, as reported by Kwak *et al.*, the creation of arcs among accounts is more complex than our simple approximation [15]. However, our goal is to understand the evolution of the macrostructure of the Twitter social graph with time, not the fine grain evolution of arcs between accounts. For this reason, we believe that our approximation is reasonable.

Moreover, to validate this approximation on creation dates of arcs, we compare our approximation with two datasets collected in 2009 [16, 9]. Kwak *et al.* [16] and Cha *et al.* [9] independently collected two Twitter datasets in August 2009 and September 2009 respectively and used different methodology. Kwak *et al.* used a technique close to

[3]For a given account, Meeder *et al.* observed that the 1.0 Twitter API returned the arcs in an order that was the reverse order of creation of the arcs for this account [19]. Our recent experiments with the Twitter API have shown that it is no more possible to rely on this ordering property.

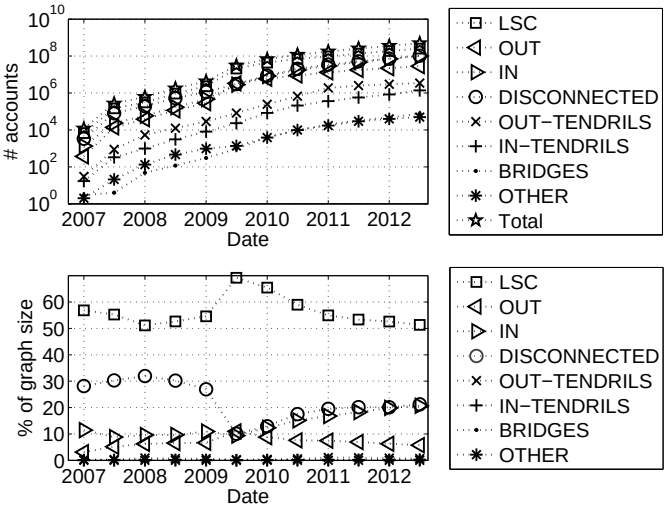

Figure 9: The estimated evolution of the macrostructure of Twitter with time. *(top) Sizes of components in log scale. (bottom) Sizes of the same components as a percentage of the size of the graph.*

a BFS and reverse BFS from a popular account and also collected accounts referring to trending topics (so active accounts only), and Cha *et al.* used a crawl by account ID (as we did). For each of the two datasets we computed the Twitter macrostructure according to the methodology described in Section 3, and we approximated the macrostructure of Twitter using our dataset in July 2009 and January 2010. We show in Figure 8 the result of this validation: the order of the size of each components is consistent between the two validation datasets and our dataset. In addition, we have compared the dataset by Kwak *et al.* with our closest estimation (July 2009). We found that 88.25% of the users common to both datasets belong to the same components in both datasets. We cannot make such a validation for the second dataset because Cha *et al.* have anonymized it. In summary, the dynamics of the creation and deletion of arcs is complex [15], but we have shown that our simple approximation is reliable enough for the purpose of our macrostructure study.

There is no DISCONNECTED component in Figure 8, because this component is missing in the two validation datasets. Either the methodology did not permit to crawl accounts in this component [16], or these accounts were filtered out in the published dataset [9]. We observe in Figure 8 some small variations for the OUT, IN-TENDRILS, and OUT-TENDRILS components between the two validation datasets and our dataset. These variations can be explained by a major change in the Twitter macrostructure that happened in 2009. We discuss further this change in the next section.

5.2 Evolution of the Macrostructure

To observe the evolution of the Twitter social graph with time, we approximate its macrostructure using the simple methodology discussed in Section 5.1 every six months from January 1, 2007 to July 1, 2012. The first account on Twitter was created on March 21, 2006, but due to the small number of accounts created between March and July 2006, we decided to skip the macrostructure of the Twitter social graph in July 2006 and start our analysis in January 2007.

We see in Figure 9 (top) the evolution of the size of each component with time, confirming that the LSC, OUT, IN, and DISCONNECTED have always been the largest components in Twitter. However, by looking at the size of each component normalized with the graph size in Figure 9 (bottom), we observe an interesting change in proportion of macrostructure components in 2009.

Before 2009, the proportion of the DISCONNECTED component was around 30%, the IN component was stable in size, and the size of the OUT component was increasing. The real public adoption of Twitter started in 2009 where the total number of accounts went from 4.265 million in January 2009 to 67.487 million in January 2010. Several events contributed to attract new users on Twitter during that period: the terrorist attacks in Mumbai was one of the first event followed on Twitter in November 2008, attracting the attention of other news media such as CNN; some influential celebrities started to use Twitter such as Oprah Winfrey, and, for the first time, some accounts reached one million followers.

We see in Figure 9 (bottom) that the large adoption of Twitter in 2009 led to changes in the macrostructure of its social graph. The proportion of the DISCONNECTED component dropped to 10% while the LSC jumped to 70%. We have seen in Section 4 that the DISCONNECTED component corresponds to abandoned accounts, so during such a large adoption phase, the proportion of abandoned accounts is much lower. However, this proportion increased in 2010 and 2011 to reach a stable value, with the DISCONNECTED component representing around 20% of all accounts.

We also observe in Figure 9 (bottom) that the proportion of the OUT component has been decreasing since 2009. The reason is that a large fraction of celebrities joined Twitter in 2009 and 2010. Some of these celebrities created an account to increase their visibility, but never intended to follow other accounts, thus they joined the OUT component. The fraction of such celebrities is decreasing compared to regular accounts, and also joining Twitter without following anyone in the LSC component is a decreasing trend. Indeed, the proportion of the IN component has been increasing since 2009, showing that it is an increasing trend to follow accounts in the LSC component without tweeting and being followed.

It is worth noticing that the two most popular Twitter datasets [16, 9] have been collected in 2009. We have seen that the Twitter social graph macrostructure has significantly changed during the period 2009/2010, calling for a newer dataset such as the one we collected, which is more representative of the actual Twitter social graph. We also note that the two datasets of 2009 do not contain accounts belonging to the DISCONNECTED component, unlike our dataset, which is an issue for researchers focusing on malicious activities and abandoned accounts on Twitter.

5.3 Distribution of New Accounts in Components

In this section, we evaluate to which component the new accounts created during each 6 months period belong to. To find this distribution, we use the approximations of the Twitter social graph macrostructure described in Sec-

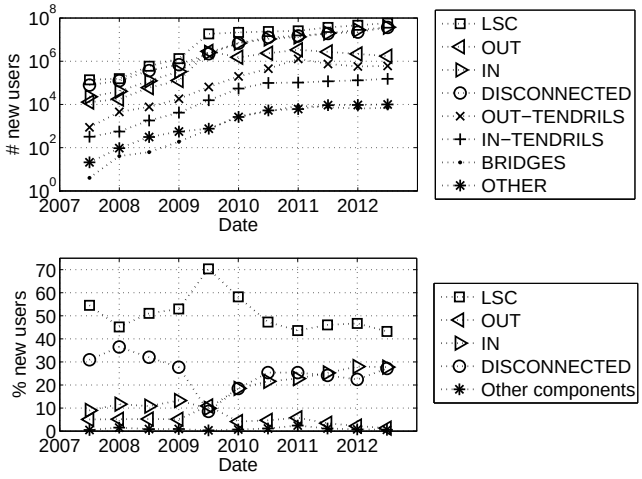

Figure 10: Distribution of new accounts per components with time. *(top) Number of new accounts per component. (bottom) Fraction of the total number of new accounts per component.*

tion 5.2. Then, for each pair of contiguous approximations in time (e.g., July 2008 and January 2009), we remove all accounts already present in the oldest one to the newest one. This way, we obtain the evolution with time of the distribution of the new accounts in components, see Figure 10.

We observe in Figure 10 (top) that the total number of new accounts increases with time for the LSC, IN, and DISCONNECTED components, but not for OUT. This decrease confirms our discussion in Section 4.2 on the OUT component, explaining that new selfish accounts are decreasing in Twitter.

Figure 10 (bottom) shows the fraction of the total number of new accounts per component. We observe that new accounts join most the LSC component, but this is a decreasing trend at the benefit of the IN and DISCONNECTED components. We explain this trend by two changes in the usage of Twitter initiated in 2009. First, passive followers are taking an increasing role in Twitter; passive followers are accounts that follow other accounts, but that are not followed and never publish tweets, as described in Section 4.3. This increasing role of passive followers shows that Twitter is more and more used as a regular information media in which people receive information, but do not produce any. However, more than 40% of new accounts are still joining the LSC component, making Twitter the largest and most participative information media. Second, as Twitter is very popular, it is attracting a large fraction of users that are just creating a Twitter account out of curiosity, but never effectively use it. Most of these accounts end up in the DISCONNECTED component.

6. RELATED WORK

Twitter has been widely studied for years. A large fraction of the literature is on the identification of malicious behavior on Twitter [24, 27, 10], on the study of tweets propagation [21, 26], and on privacy [12, 18]. All these studies are not directly related to our work as they do not crawl the Twitter social graph and do not explore its properties.

Closer to our work, several studies focused on the Twitter social graph. Some of them crawled partially the graph before 2009 [13, 14, 11], so before the wide adoption of Twitter. Two studies made a large crawl of the Twitter social graph. Kwak *et al.* used a technique close to a BFS and reverse BFS from a popular account and also collected accounts referring to trending topics. This crawling methodology cannot capture some users that are not connected to the LSC component, and that do not tweet about trending topic, thus a partial view of the Twitter social graph. Cha *et al.* [9] used a crawl by account ID, that is close to what we did. Both of these studies made their dataset publicly available and others built on it [17, 25, 8, 22], but the datasets were collected in 2009 during the main change in the Twitter social graph we discussed in Section 5.2.

To the best of our knowledge, the dataset we present is the most up-to-date and the most complete description of the Twitter social graph. Moreover, none of these studies explores the macrostructure of the Twitter social graph, a new way to represent directed social graphs. Broder *et al.* [7] introduced first the notion of macrostructure for a directed graph in the context of the Web, but we significantly improved it, and we are the first ones to apply it to Twitter. Unlike what Broder *et al.* proposed, we present a methodology to compute the exhaustive macrostructure of any large directed social graph, along with the categorization of each account in the identified component, which is a significant methodological step.

7. CONCLUSION

In this paper, we present the largest, most complete, and most up-to-date crawl of the Twitter social graph. This graph contains 505 million accounts connected with 23 billion arcs. In addition, we present a methodology to practically compute the macrostructure of any directed social graph and to exhaustively classify each account to one of the identified components. We applied this methodology to the Twitter social graph and found that only 50.71% of the accounts belong to the LSC component, and that 21.60% of the accounts (in the DISCONNECTED component) have no path to the other accounts.

We show that the main components of the macrostructure of the Twitter social graph correspond to specific usages. For instance, the LSC component hold most of the regular Twitter activity, and the IN component holds passive followers. Finally, we present a simple methodology to explore the evolution of the macrostructure of Twitter with time, we validate this methodology, and we show that the public datasets crawled in 2009 do not represent the current macrostructure of the Twitter social graph.

We believe that our collected dataset is a gold mine for any researcher working on social graphs and that the macrostructure analysis sheds a new light on the Twitter social graph that will be useful for both theoreticians and experimenters.

8. ACKNOWLEDGEMENTS

The authors thank Krishna P. Gummadi (MPI-SWS) for insights on the analysis of the dataset we collected and valuable feedback from the early stages of this work. We also thank him for sharing the list of influential Twitter users identified by Sharma *et al.* [22].

9. REFERENCES

[1] soTweet: Studying Twitter at Scale. http://www-sop.inria.fr/members/Arnaud.Legout/Projects/sotweet.html

[2] PlanetLab. https://www.planet-lab.org/

[3] Twitter REST API 1.0. https://dev.twitter.com/docs/api/1

[4] FAQs about verified accounts. https://support.twitter.com/groups/31-twitter-basics/topics/111-features/articles/119135-about-verified-accounts

[5] About public and protected Tweets. https://support.twitter.com/entries/14016

[6] How to deactivate your account. https://support.twitter.com/articles/15358-how-to-deactivate-your-account

[7] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. In *Proc. of WWW'00*, Amsterdam, The Netherlands, April 2000.

[8] M. Cha, F. Benevenuto, H. Haddadi, and K. P. Gummadi. The world of connections and information flow in Twitter. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions*, 42(4):991–998, 2012.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proc. of AAAI ICWSM'10*, Washington DC, USA, May 2010.

[10] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. Understanding and combating link farming in the Twitter social network. In *Proc. of WWW'12*, Lyon, France, April 2012.

[11] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1), 2008.

[12] L. Humphreys, P. Gill, and B. Krishnamurthy. How much is too much? Privacy issues on Twitter. In *Proc. of the Conference of International Communication Association*, Singapore, June 2010.

[13] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: understanding microblogging usage and communities. In *Proc. of WebKDD/SNA-KDD'07*, San Jose, California, August 2007.

[14] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about Twitter. In *Proc. of WOSN'08*, Seattle, WA, USA, August 2008.

[15] H. Kwak, H. Chun, and S. Moon. Fragile online relationship: a first look at unfollow dynamics in Twitter. In *Proc. of ACM CHI'11*, Vancouver, BC, Canada, 2011.

[16] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. of WWW'10*, Raleigh, NC, USA, May 2010.

[17] J. G. Lee, P. Antoniadis, and K. Salamatian. Faving reciprocity in content sharing communities: A comparative analysis of Flickr and Twitter. In *Proc. of ASONAM'10*, Odense, Denmark, August 2010.

[18] H. Mao, X. Shuai, and A. Kapadia. Loose tweets: an analysis of privacy leaks on Twitter. In *Proc. of WPES'11*, Chicago, IL, USA, October 2011.

[19] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. We know who you followed last summer: inferring social link creation times in twitter. In *Proc. of WWW'11*, Hyderabad, India, March 2011.

[20] M. Russell. *21 Recipes for Mining Twitter*. Real Time Bks. O'Reilly Media, Inc., 2011.

[21] E. Sadikov and M. M. M. Martinez. Information propagation on Twitter. CS322 project report, Stanford University, 2009.

[22] N. K. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. P. Gummadi. Inferring who-is-who in the Twitter social network. In *Proc. of ACM WOSN'12*, Helsinki, Finland, August 2012.

[23] R. Tarjan. Depth-first search and linear graph algorithms. In *Proc. of 12th Annual Symposium on Switching and Automata Theory*, 1971.

[24] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: an analysis of Twitter spam. In *Proc. of ACM SIGCOMM IMC'11*, Berlin, Germany, November 2011.

[25] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on Twitter. In *Proc. of WWW'11*, Hyderabad, India, March 2011.

[26] S. Ye and S. F. Wu. Measuring message propagation and social influence on Twitter.com. In *Proc. of SocInfo'10*, Laxenburg, Austria, October 2010.

[27] C. M. Zhang and V. Paxson. Detecting and analyzing automated activity on Twitter. In *Proc. of PAM'11*, Atlanta, GA, USA, March 2011.